Chapter IV
Using Biomedical Terminological Resources for Information Retrieval

Piotr Pezik
European Bioinformatics Institute, Wellcome Trust Genome Campus, UK

Antonio Jimeno Yepes
European Bioinformatics Institute, Wellcome Trust Genome Campus, UK

Dietrich Rebholz-Schuhmann
European Bioinformatics Institute, Wellcome Trust Genome Campus, UK

ABSTRACT

The present chapter discusses the use of terminological resources for Information Retrieval in the biomedical domain. The authors first introduce a number of example resources which can be used to compile terminologies for biomedical IR and explain some of the common problems with such resources including redundancy, term ambiguity, and insufficient coverage of concepts and incomplete Semantic organization of such resources for text mining purposes. They also discuss some techniques used to address each of these deficiencies, such as static polysemy detection as well as adding terms and linguistic annotation from the running text. In the second part of the chapter, the authors show how query expansion based on using synonyms of the original query terms derived from terminological resources potentially increases the recall of IR systems. Special care is needed to prevent a query drift produced by the usage of the added terms and high quality word sense disambiguation algorithms can be used to allow more conservative query expansion. In addition, they present solutions that help focus on the user's specific information need by navigating and rearranging the retrieved documents. Finally, they explain the advantages of applying terminological and Semantic resources at indexing time. The authors

argue that by creating a Semantic index with terms disambiguated for their Semantic types and larger chunks of text denoting entities and relations between them, they can facilitate query expansion, reduce the need for query refinement and increase the overall performance of Information Retrieval. Semantic indexing also provides support for generic queries for concept categories, such as genes or diseases, rather than singular keywords.

INTRODUCTION

Researchers in the life science domain have high hopes for the automatic processing of scientific literature. Consequently, there has been a growing interest in developing systems retrieving domain-specific information without making users read every document. (Rebholz-Schuhmann et al., 2005). Over the recent years, Text Mining and Information Retrieval in the life science domain have evolved into a specialized research topic, as the exploitation of biomedical data resources becomes more and more important (Hirschman et al., 2005/2008).

The genomics era has lead to the generation of large-scale data resources containing sequence information about genes and proteins (EMBL database, UniProtKb), and keep track of experimental findings such as gene expression profiles measured in MicroArray experiments (GEO, ArrayExpress). All of these types of data require specialized databases that represent scientific findings with the highest level of detail possible. To address this need, standardization efforts have been launched to develop well-defined database schemas and controlled vocabularies that represent named concepts and entities, for example genes, proteins, parts of genes, functions of proteins and the biological processes that proteins and chemicals are involved in.

In principle, text mining solutions (both information retrieval as well as information extraction oriented ones) benefit from the availability of terminological resources. In the case of information retrieval such resources improve a number of tasks, such as the expansion of user queries, and recognition of concepts in texts, linking them to existing databases and conducting a first order analysis of textual data. Terminological resources are essential in tackling the problem of synonymy, where a single concept has multiple orthographic representations as well as that of polysemy, where a single term may refer to multiple concepts. Lexical metadata encoding hypernymic relations between terms may also come in handy when implementing search engines supporting queries for a generic Semantic type rather than for a specific keyword.

In this chapter we first focus on the availability of terminological resources for the biomedical domain, and discuss different aspects of adopting them for Information Retrieval tasks (section 2 and 3). Then we introduce query refinement as one use scenario of lexicons for improved Information Retrieval (section 4). Finally, we present Semantic indexing as an alternative and complementary approach to integrating the use of terminological resources into the early stages of text processing of biomedical literature (section 5).

COMPILATION OF LEXICAL RESOURCES

A number of life science data resources lending support to text mining solutions are available, although they differ in quality, coverage and suitability for IR solutions. In the following section we provide an outline of the databases commonly used to aid Information Retrieval and Information Extraction.